Asses node importance: Centrality measures

NODE

- We can measure nodes importance using so-called **centrality**.
- Bad term: nothing to do with being central in general
- Usage:
  - Some centralities have straightforward interpretation
  - Centralities can be used as *node features* for machine learning on graph
    - (Classification, link prediction, …)

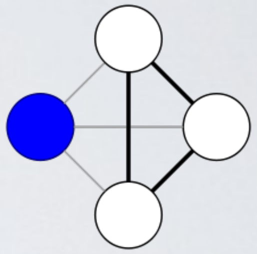
Connectivity based centrality measures

NODE DEGREE

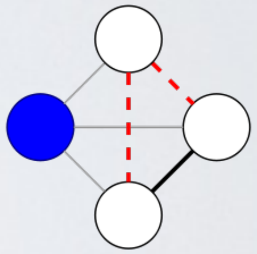
- **Degree**: how many neighbors
- Often enough to find important nodes
  - Main characters of a series talk with the more people
  - Largest airports have the most connections
  - …
- But not always
  - Facebook users with the most friends are spam
  - Webpages/wikipedia pages with most links are simple lists of references
  - …

NODE CLUSTERING COEFFICIENT

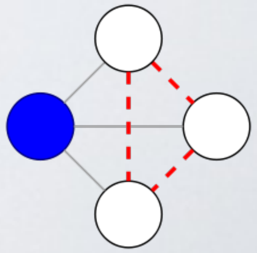
- **Clustering coefficient**: closed triangles/triads
- Tells you if the neighbors of the node are connected
- Be careful!
  - Degree 2: value 0 or 1
  - Degree 1000: Not 0 or 1 (usually)
  - Ranking them is not meaningful
- Can be used as a proxy for "communities" belonging:
  - If node belong to single group: high CC
  - If node belong to several groups: lower CC



c = 1



c = 1/3



c = 0

RECURSIVE DEFINITIONS

- Recursive importance:
  - **Important nodes** are those connected ***to important nodes***
- Several centralities based on this idea:
  - Eigenvector centrality
  - PageRank
  - Katz centrality
  - …

RECURSIVE DEFINITION

- We would like scores such as :
  - Each node has a score (centrality),
  - If every node "sends" its score to its neighbors, the sum of all scores received by each node will be equal to its original score

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

x_i is the centrality of node i.

$A_{ij} = 1$ if there is an edge, 0 otherwise

RECURSIVE DEFINITION

- This problem can be solved by what is called the *power method:*
  - 1) We initialize all scores to random values
  - 2)Each score is updated according to the desired rule, until reaching a stable point (after normalization)

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)}$$

- Why does it converge?
  - Perron-Frobenius theorem for *real* and *irreducible square matrices* with *non-negative* entries
  - =>True for undirected graphs with a single connected component

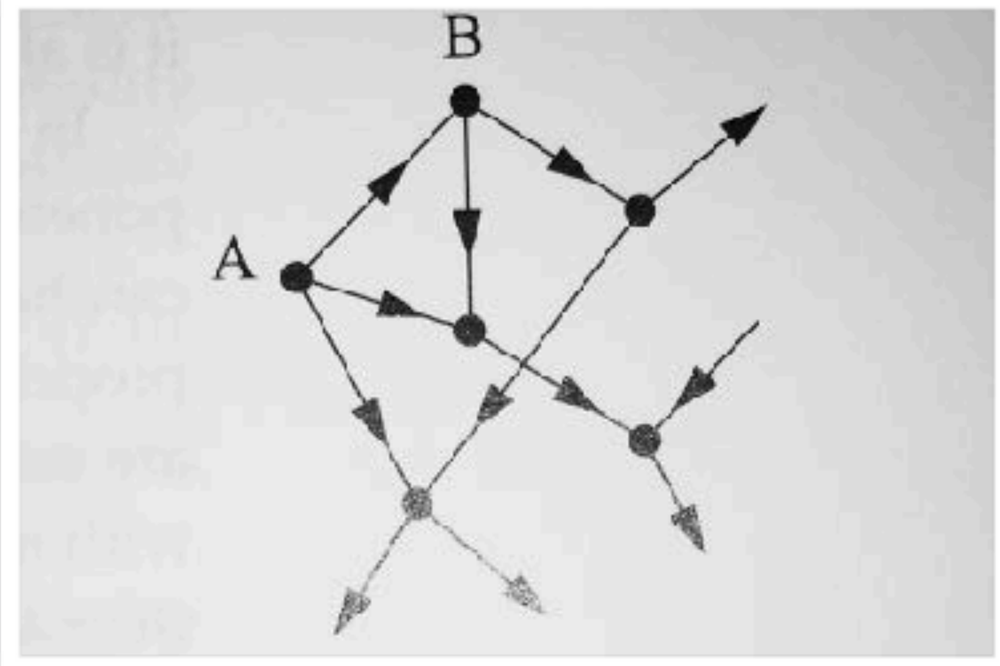
EIGENVECTOR CENTRALITY

- What we just described is called the Eigenvector centrality
- A couple eigenvector (x) and eigenvalue (λ) is defined by the following relation: $Ax = \lambda x$
  - x is a vector of size *n*, which can be interpreted as the scores of nodes
  - Ax yield a new vector of size *n*, which corresponds for each node to receive the sum of the scores of its neighbors (like in the power method)
  - The equality means that the new scores are proportional to the previous scores
- What Perron-Frobenius algorithm says is that the power method will always converge to the *leading eigenvector*, i.e., the eigenvector associated with the highest eigenvalue

Eigenvector Centrality

Some problems in case of directed network:

- Adjacency matrix is asymmetric
- 2 sets of eigenvectors (Left & Right)
- 2 leading eigenvectors
  - Use right eigenvectors : consider nodes that are pointing towards you



But problem with source nodes (0 in-degree)

-Vertex A is connected but has only outgoing link = Its centrality will be 0

-Vertex B has outgoing and an incoming link, but incoming link comes from A = Its centrality will be 0

-etc.

Solution: Only in strongly connected component

Note: Acyclic networks (citation network) do not have strongly connected component

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

PageRank Centrality

(Side notes)

-"We choose our system name, Google, because it
is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search "

-"[…] at the same time, search engines have migrated from the academic domain to the commercial. **Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong goal to push more development and understanding into the academic realm."**

-"[...], **we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers**."

PAGERANK

- 2 main improvements over eigenvector centrality:
  - In directed networks, problem of source nodes
    - => Add a constant centrality gain for every node
  - Nodes with very high centralities give very high centralities to all their neighbors (even if that is their only in-coming link)
    - => What each node "is worth" is divided equally among its neighbors (normalization by the degree)

$$x_i^{(t+1)} = \sum_{j=1}^{n} A_{ij} x_j^{(t)} \qquad => \qquad x_i = \alpha \sum_{j=1}^{n} A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta$$

With by convention β=1 and α a parameter (usually 0.85)

PageRank - as Random Walk

Main idea: The PageRank computation can be interpreted as a Random Walk process with restart

Teleportation probability: the parameter α gives the probability that in the next step of the RW will follow edges of the graph, or with probability $1-\alpha$ it will jump to a random node

- If $\alpha<1$, it assures that the RW will never be stuck at nodes with $k^{out}=0$, but it can restart the RW from a randomly selected other node

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Katz centrality of node i=

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Repeat for all distances as long
As possible (convergence)

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Sum for each node **j**

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Alpha is a parameter.
Its strength decreases at
each iteration (increased distance)

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty}\sum_{j=1}^{n}\alpha^{k}(A^{k})_{ij}$$

Number of different paths from **i** *to* **j** of length k

KATZ CENTRALITY

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ij}$$

Sum of paths to all other nodes at each distance multiplied by a factor decreasing with distance

Katz Centrality

It measures the relative degree of influence of a node within a network

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (A^k)_{ij}$$

connected pairs of nodes in distance k

attenuation factor to penalise influence by distance

- Attenuation factor α must be smaller than $1/|\lambda_0|$, i.e. the reciprocal of the absolute value of the largest eigenvalue of $\mathbf{A}$.

Matrix form:

$$\overrightarrow{C}_{Katz} = ((I - \alpha A^T)^{-1} - I)\vec{I}$$

- where I is the identity matrix, and $\vec{I}$ is the identity vector
- Katz centrality is useful for directed networks (citation nets, WWW) where Eigenvector centrality fails

Geometric centrality measures

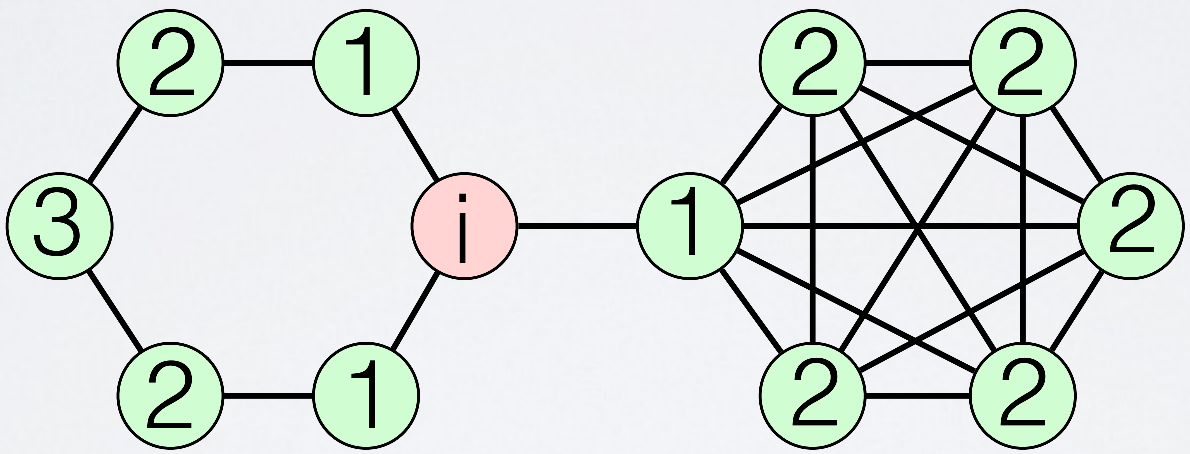
CLOSENESS CENTRALITY

$$C_{cl}(i) = \frac{n-1}{\sum_{d_{ij}<\infty} d_{ij}}$$

- **Farness**: average of length of shortest paths to all other nodes.
- **Closeness**: inverse of the Farness (normalized by number of nodes)
  - Highest closeness = More central
  - Closness=1: directly connected to all other nodes
- Well defined only on connected networks

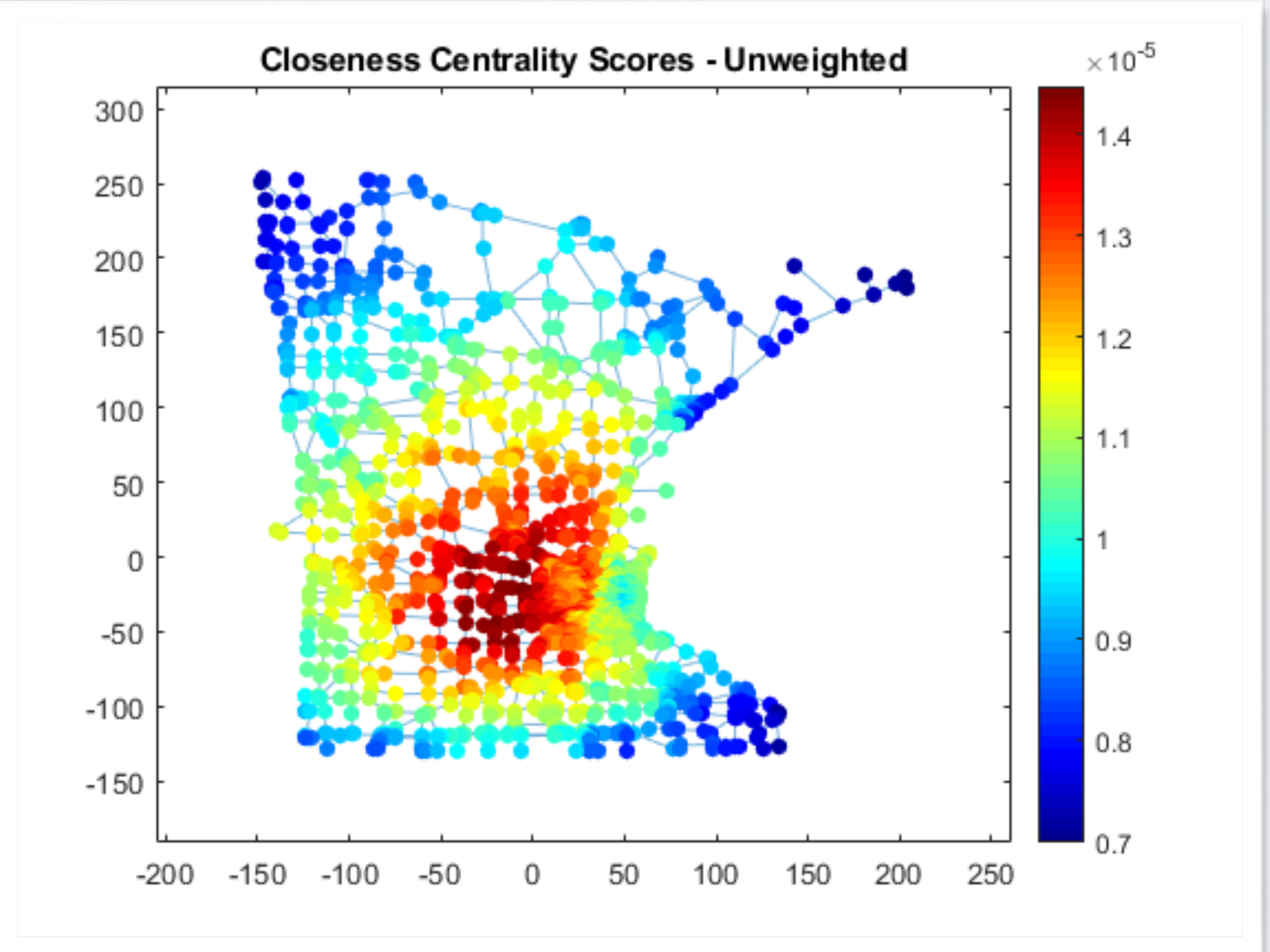
CLOSENESS CENTRALITY

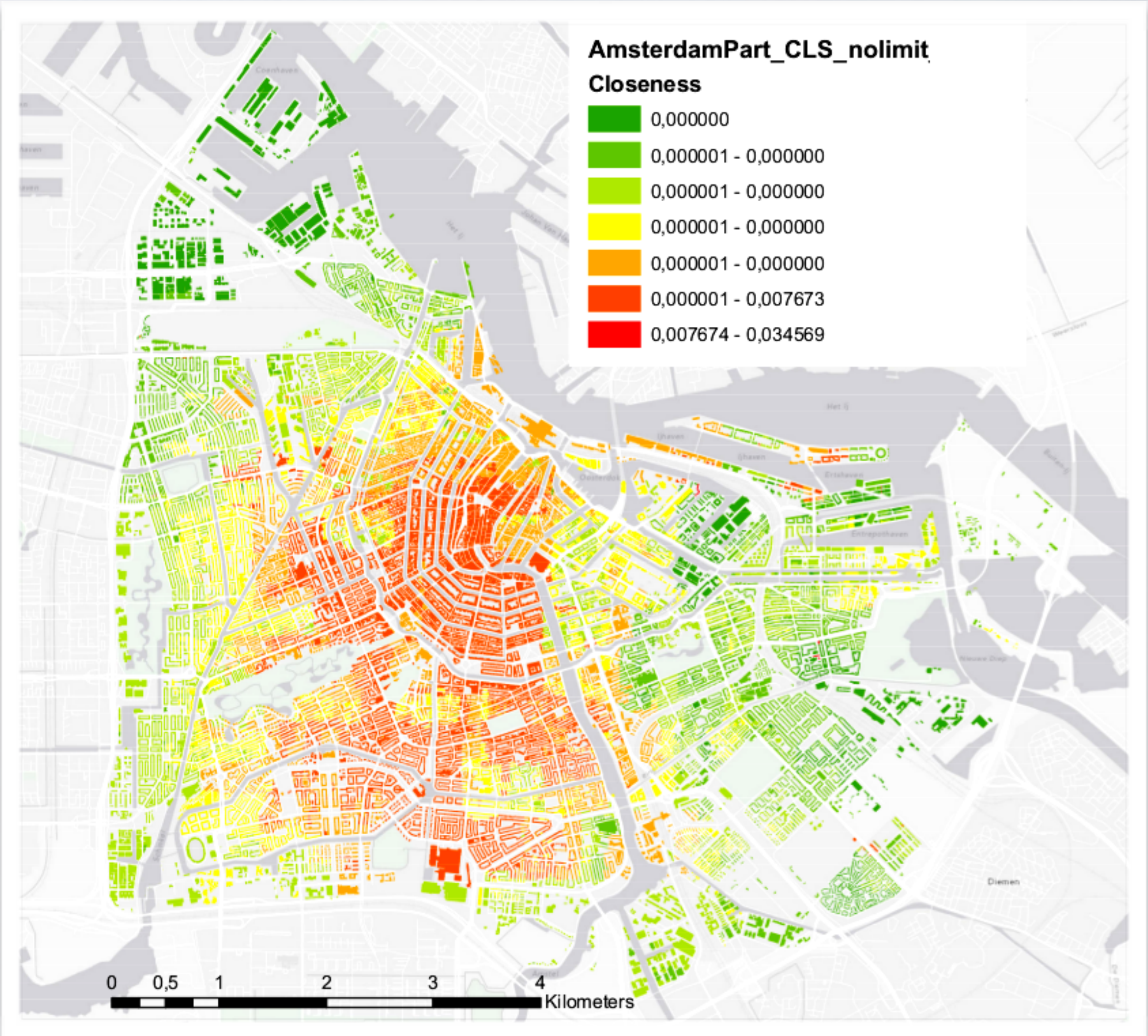
$$C_{cl}(i) = \frac{n-1}{\sum_{d_{ij}<\infty} d_{ij}}$$



$$C_{cl}(i) = \frac{12-1}{(3 \times 1 + 7 \times 2 + 1 \times 3)} = \frac{11}{20} = 0.55$$

CLOSENESS CENTRALITY





Betweenness Centrality

Assumption: important vertices are bridges over which information flows

Practically: if information spreads via shortest paths, important nodes are found on many shortest paths

Notation: $\sigma_{jk}(i)$ = number of geodesic path from j to k via i: $j \to \ldots \to i \to \ldots \to k$

σ_{jk} = number of geodesic path from j to k: $j \to \ldots \to k$

Definition:

$$C_b(i) = \sum_{j \neq k} \frac{\#\{\text{geodesic path: } j \to \ldots \to i \to \ldots \to k\}}{\#\{\text{geodesic path: } j \to \ldots \to k\}} = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

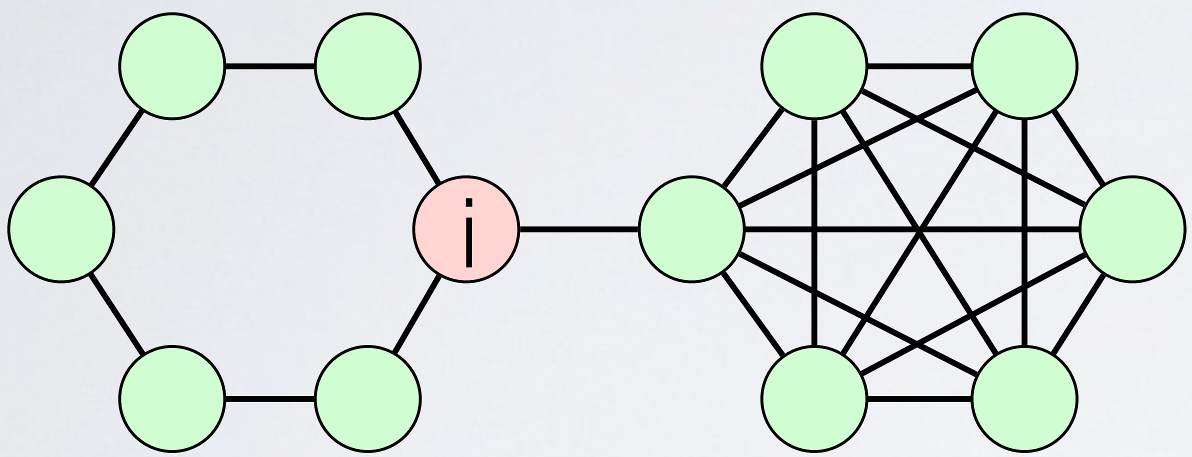
Normalised definition:

$$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$ where $$C_b \in [0,1]$$

Total number of ordered vertex pairs

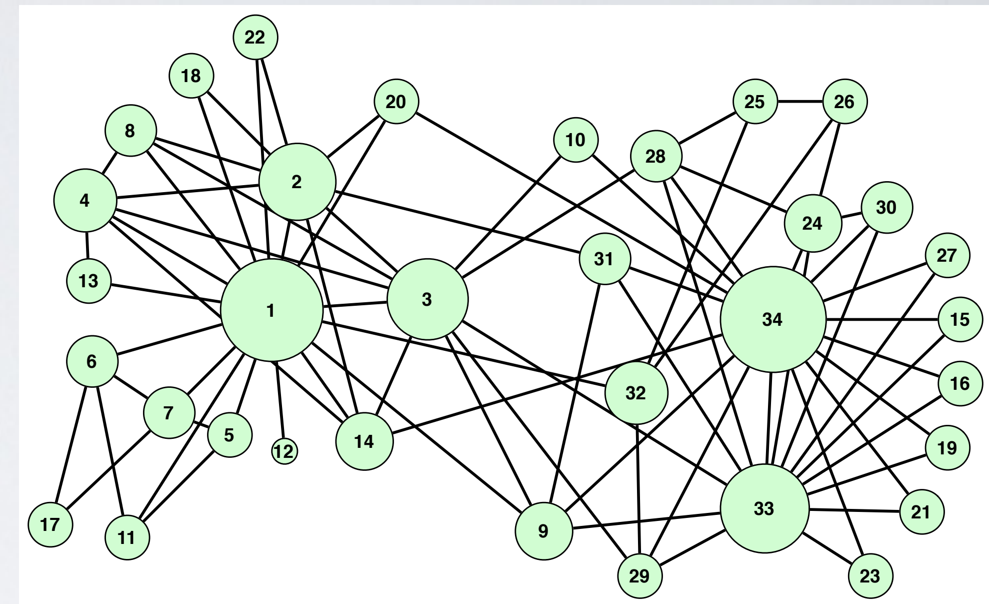
Betweenness Centrality

$$C_b(i) = \frac{1}{n^2} \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \qquad \text{where} \qquad C_b \in [0,1]$$

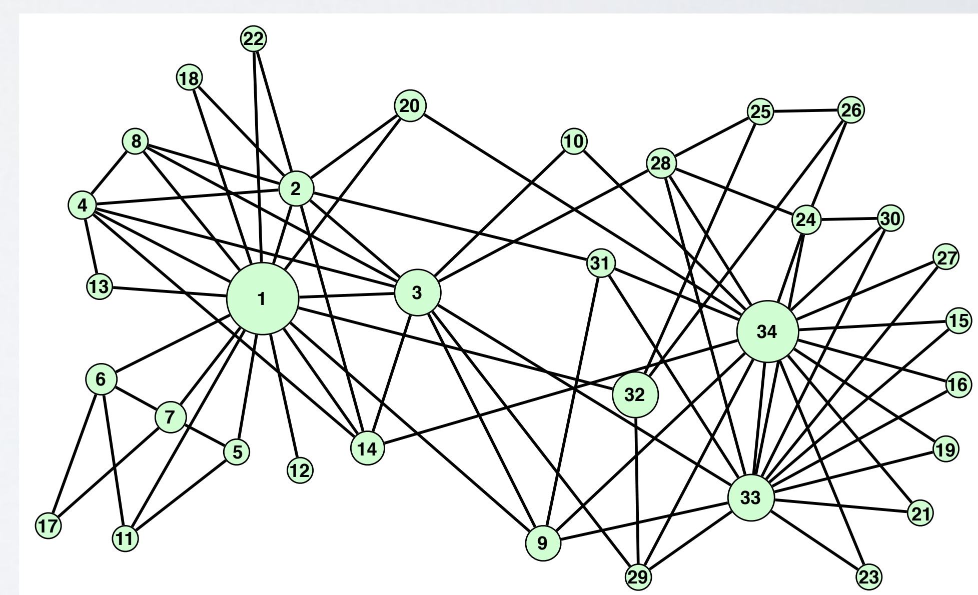


$$C_b(i) = \frac{78}{144}$$

Zachary's karate club network



degree



betweenness

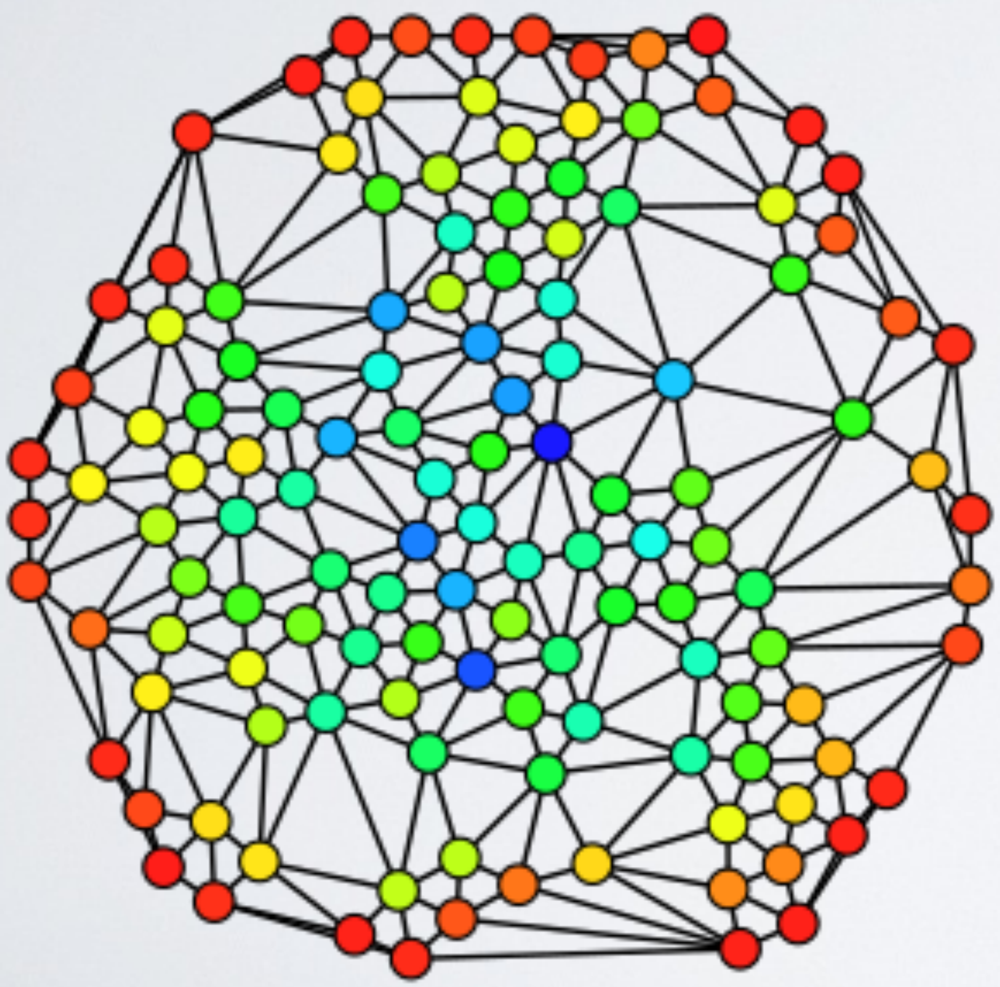
Exact computation:

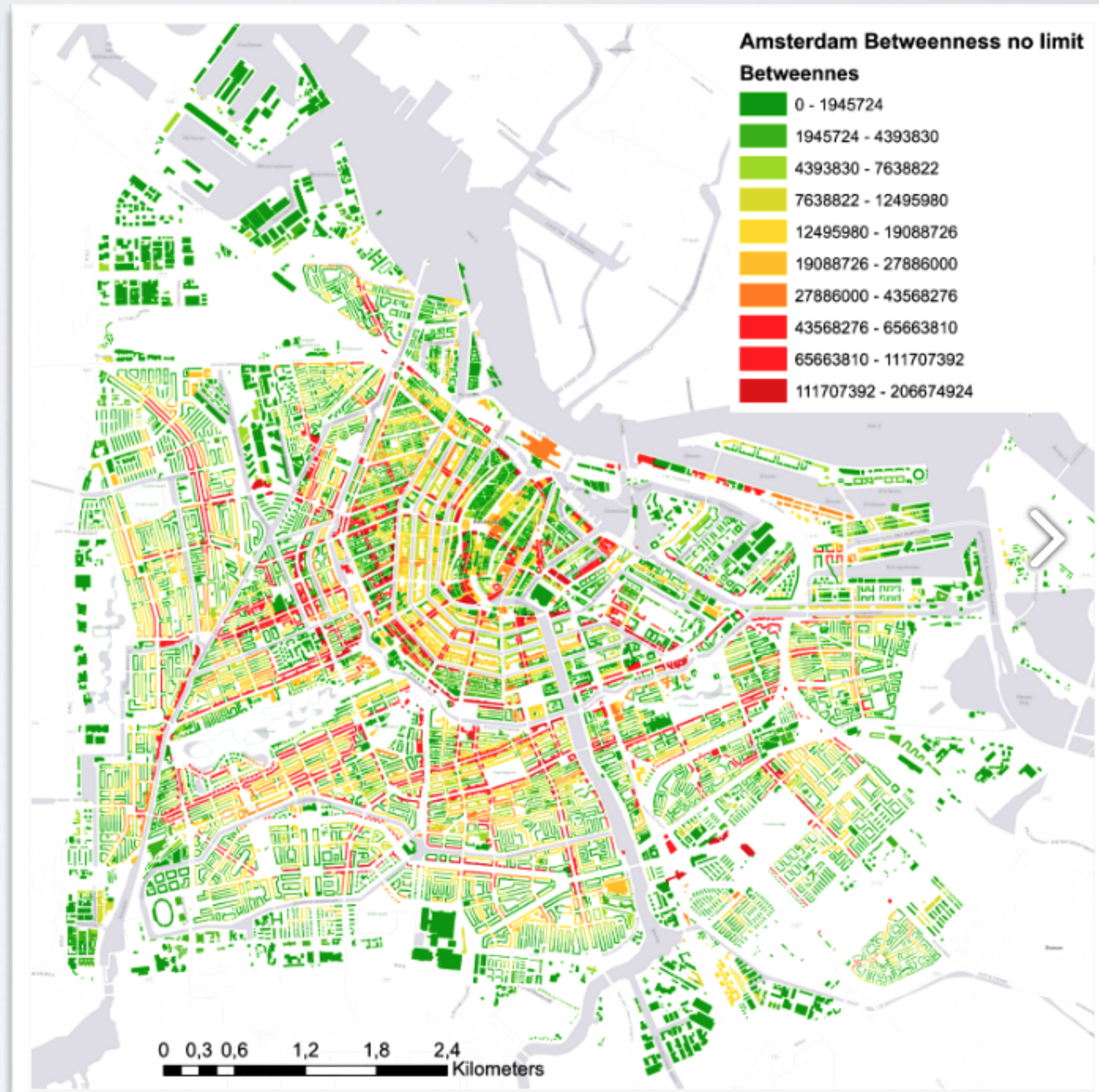
Floyd-Warshall: *O(n³) time complexity*
O(n²) space complexity

Approximate computation

Dijskstra: *O(n(m+n log n)) time complexity*

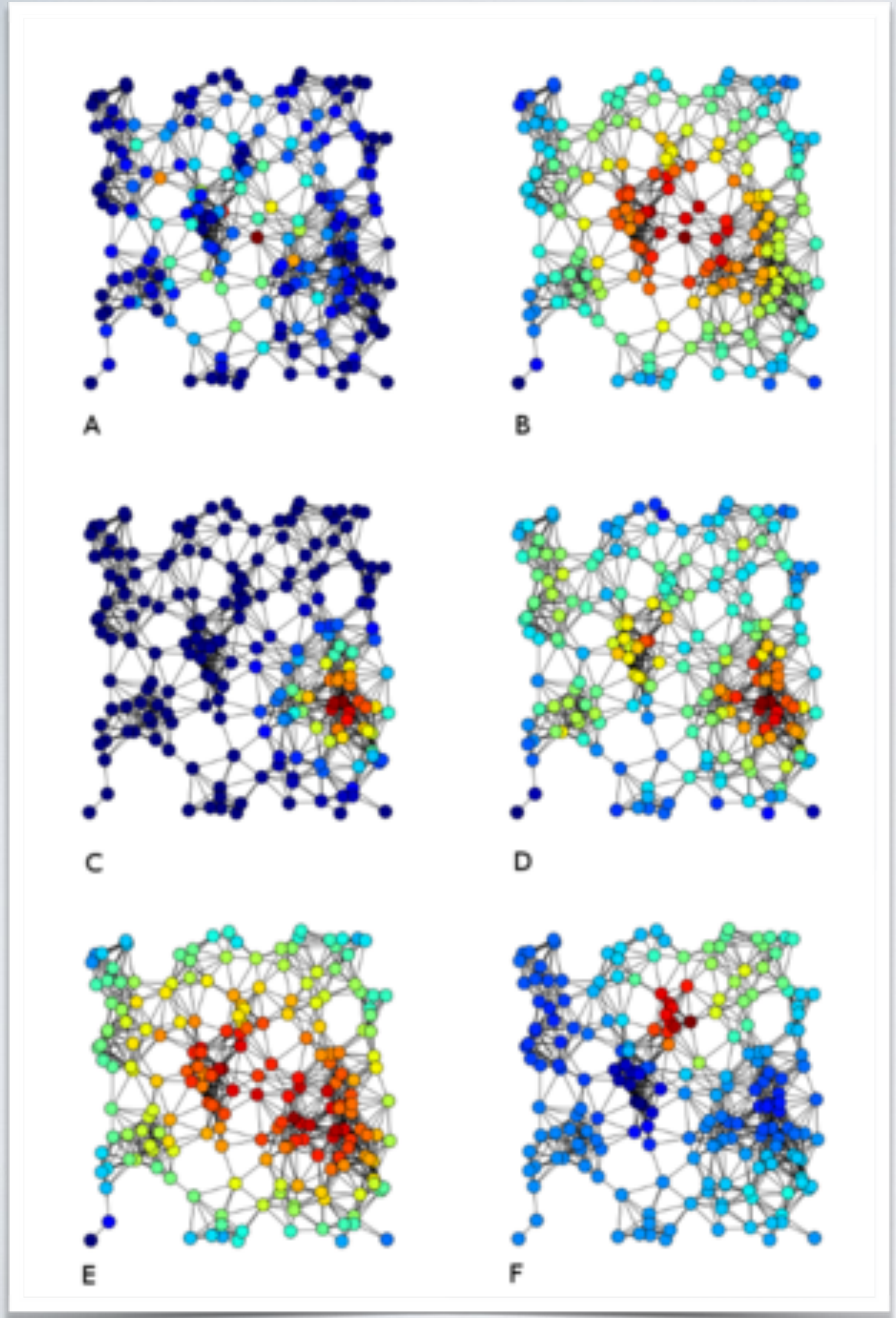
BETWEENNESS CENTRALITY





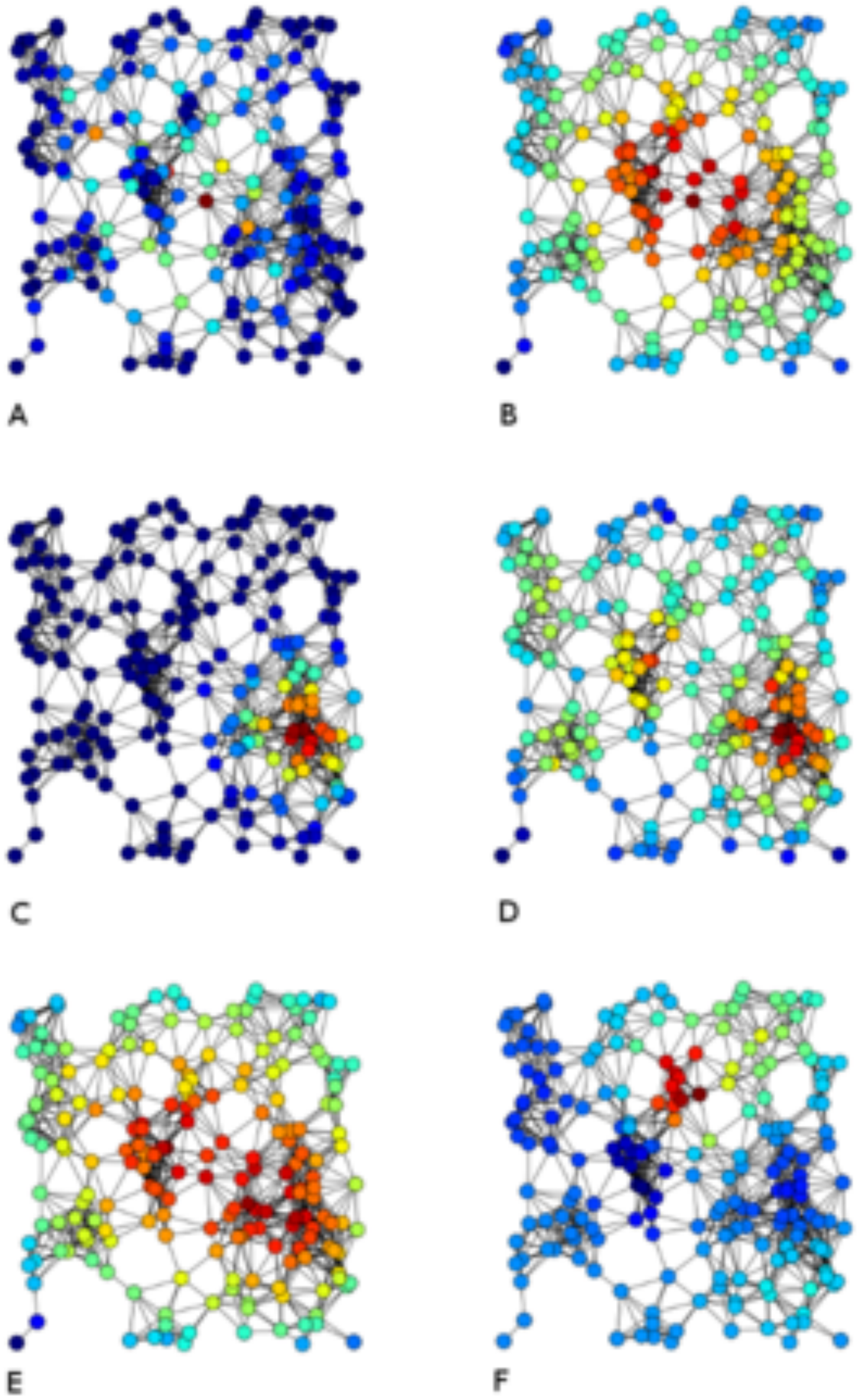
BETWEENNESS

Can you guess the node/edge of highest betweenness in the European rail network ?

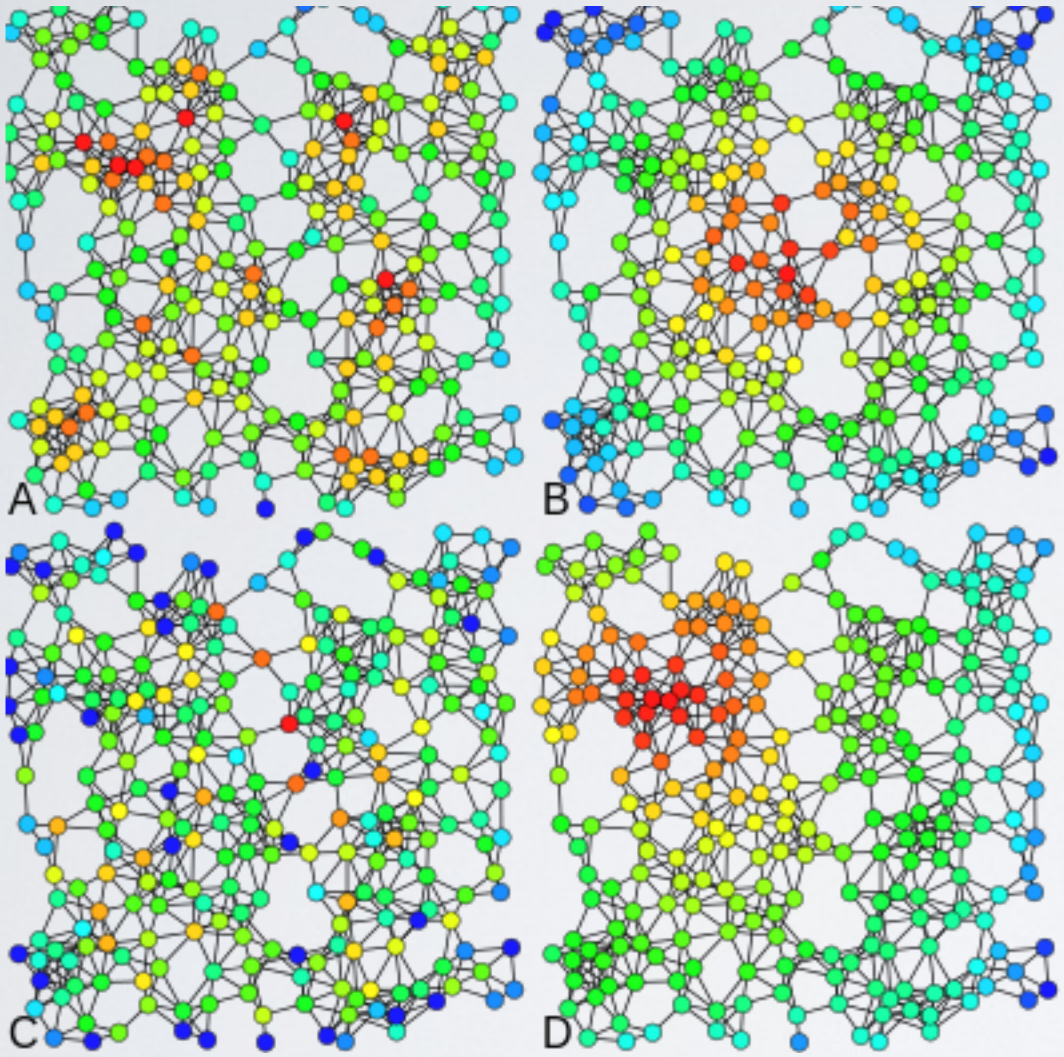


Which is which?

Harmonic
Closeness
Betweenness
Eigenvector
Katz
Degree



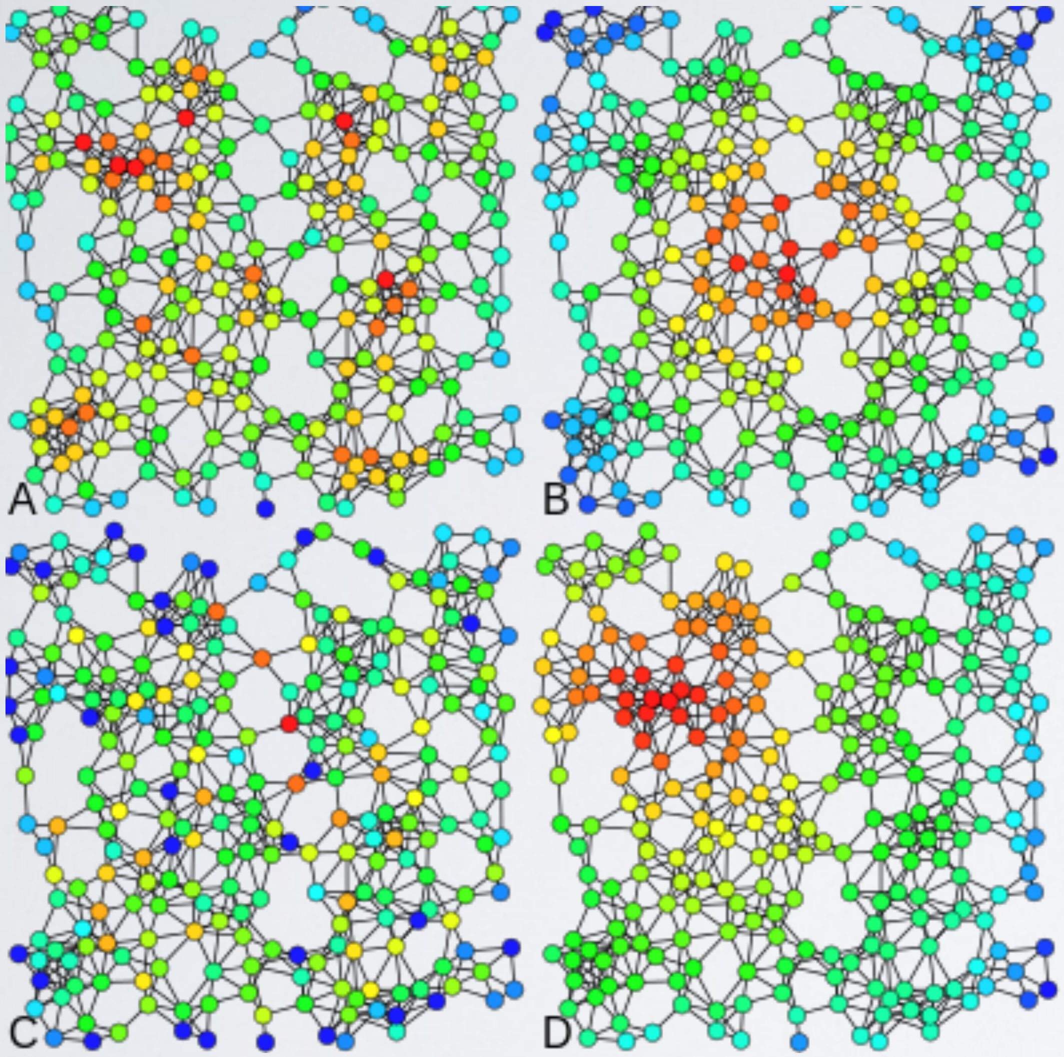
A: Betweenness
B:Closeness
C:Eigenvector
D:Degree
E:Harmonic
F: Katz



Try again :)

Degree
Betweenness
Closeness
Eigenvector

Try again :)



A: Degree
B:Closeness
C: Betweenness
D: Eigenvector

Similarity measures

Node similarity

Similarity between nodes based on their neighborhood

How much two nodes are similarly connected

- What does it mean that they have 3 neighbours in common?
- It is relative to their degree (different meaning for nodes with 3 or 100 neighbours)

➡**Normalisation to penalise nodes with small degrees**

We can define it using existing measures:

- Cosine Similarity
- Pearson Coefficient

Cosine similarity

Cosine similarity between two non-zero vectors:

$$\cos\theta = \frac{x.y}{|x||y|}$$

Number of common neighbours:

$$n_{ij} = \sum_k A_{ik}A_{kj}$$

Vectors are the rows of adjacency matrix

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}}$$

Cosine similarity: $$\boxed{\sigma_{ij} = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}}$$

Number of common neighbours normalised by the geometric mean of their degrees

Pearson coefficient

Correlation between rows of the adjacency matrix

$$r_{ij} = \frac{cov(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i \rangle)^2} \sqrt{\sum_k (A_{jk} - \langle A_j \rangle)^2}}$$

cov: covariance, expected product of deviations from individual expected values

σ: std deviation, square root of the expected squared deviation from the mean

Intuition, numerator: Number of common neighbours compared to the expected number of common neighbours

$$\sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle) = \sum_k A_{ik} A_{jk} - \frac{k_i k_j}{n}$$

Properties

- $r(i,j)=0$ - if the number of common neighbours exactly as many as we would expect by chance
- $r(i,j)>0$ - if nodes have more neighbours in common than expected
- $r(i,j)<0$ - if nodes have fewer neighbours in common than expected

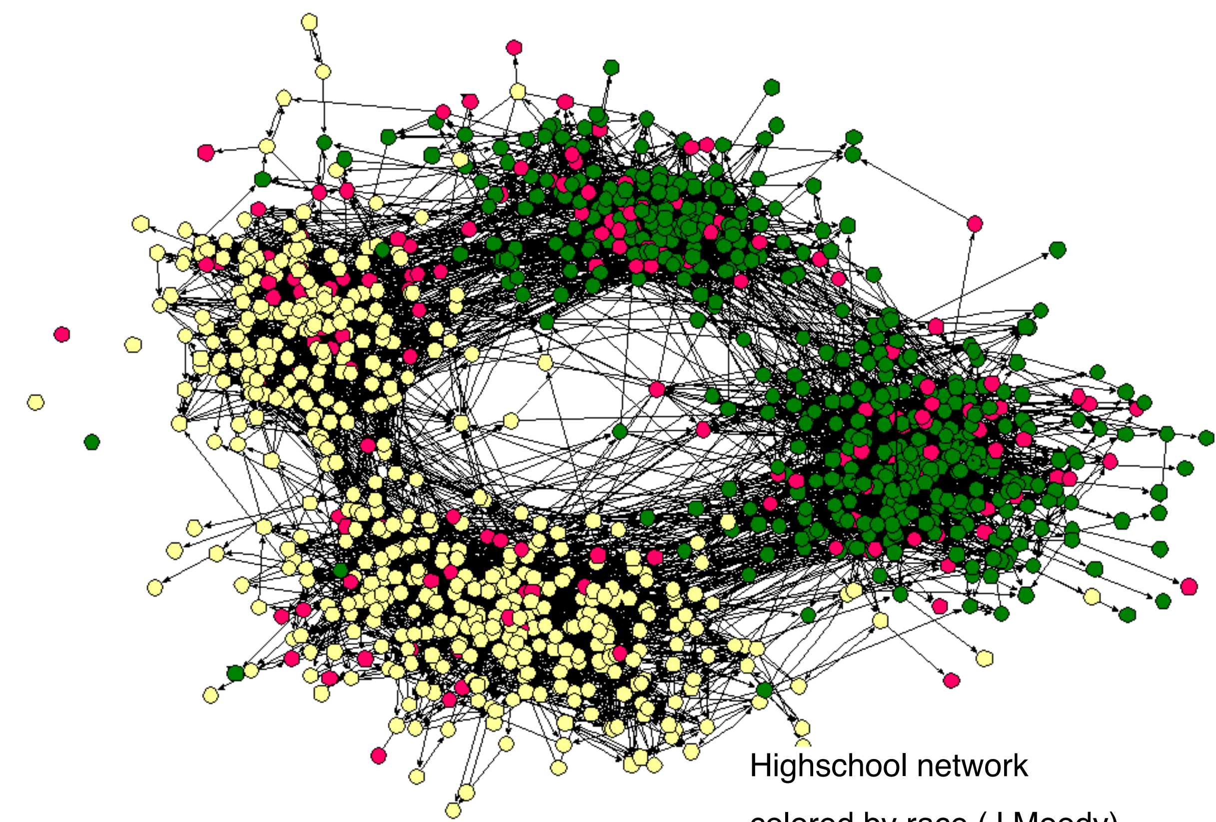
Homophily - Assortative mixing

"birds of a feather flock together"

- Property of (social) networks that nodes with similar properties tends to be connected with a higher probability than expected
- It appears as correlation between vertex properties of $x(i)$ and $x(j)$ if $(i,j)\in E$

Vertex properties

- age
- gender
- nationality
- political beliefs
- socioeconomic status
- habitual place
- obesity
- …



Highschool network
colored by race (J Moody)

- Homophily can be a link creation mechanism or consequence of social influence (and it is difficult to distinguish)

? Connected people of the same political opinion are connected because they were a priori similar (homophily) or they become similar after they become connected (social influence)?

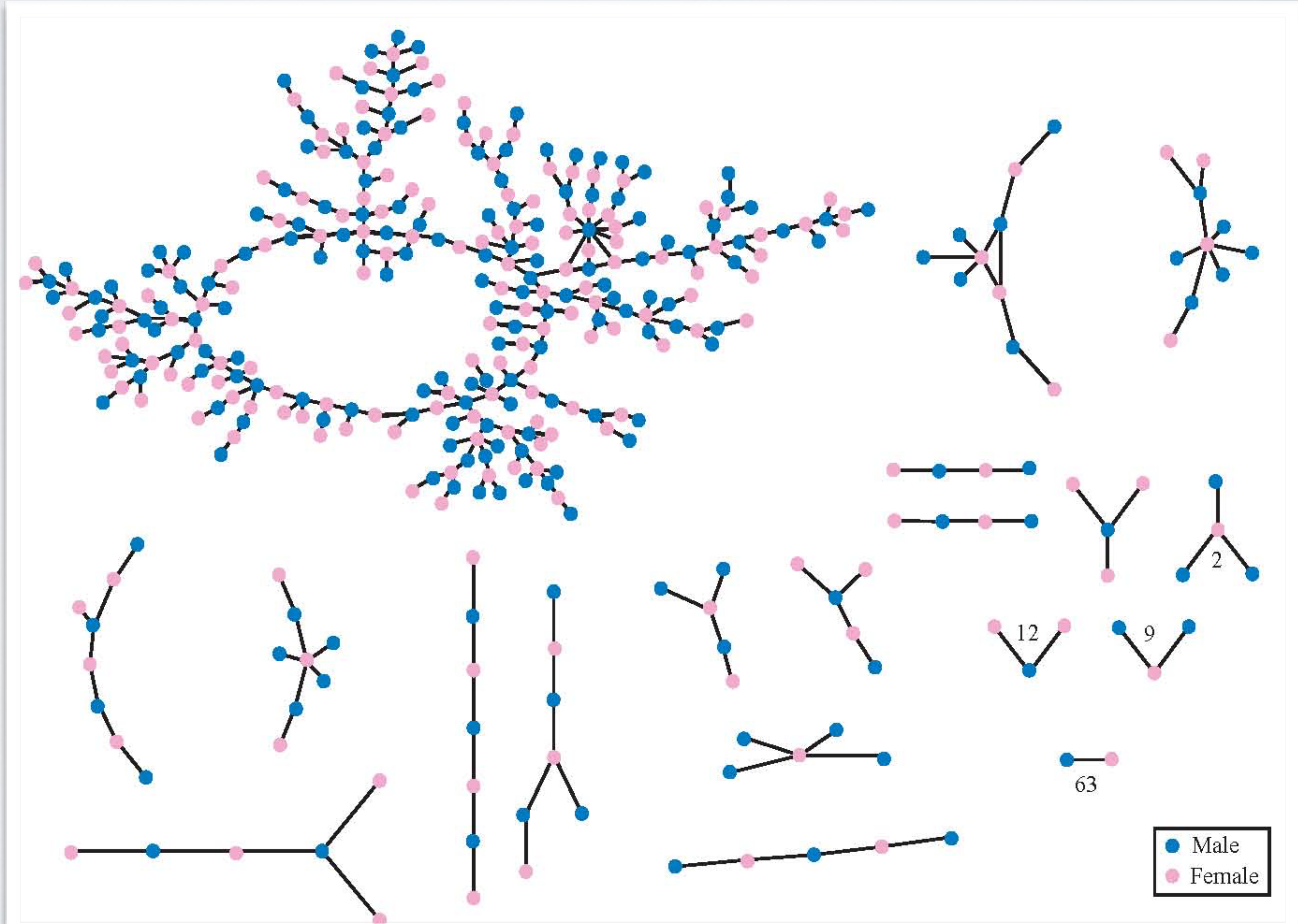
Homophily - Assortative mixing

Dissasortative mixing

- Contrary of homophily, where dissimilar nodes are tend to be connected

Examples

- Sexual networks
- Predator - prey ecological networks



Homophily - Assortative mixing

To quantify homophily

Discrete properties

		women				
		black	hispanic	white	other	a_i
men	black	0.258	0.016	0.035	0.013	0.323
	hispanic	0.012	0.157	0.058	0.019	0.247
	white	0.013	0.023	0.306	0.035	0.377
	other	0.005	0.007	0.024	0.016	0.053
	b_i	0.289	0.204	0.423	0.084	

TABLE I: The mixing matrix e_{ij} and the values of a_i and b_i for sexual partnerships in the study of Catania *et al.* [23]. After Morris [24].

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

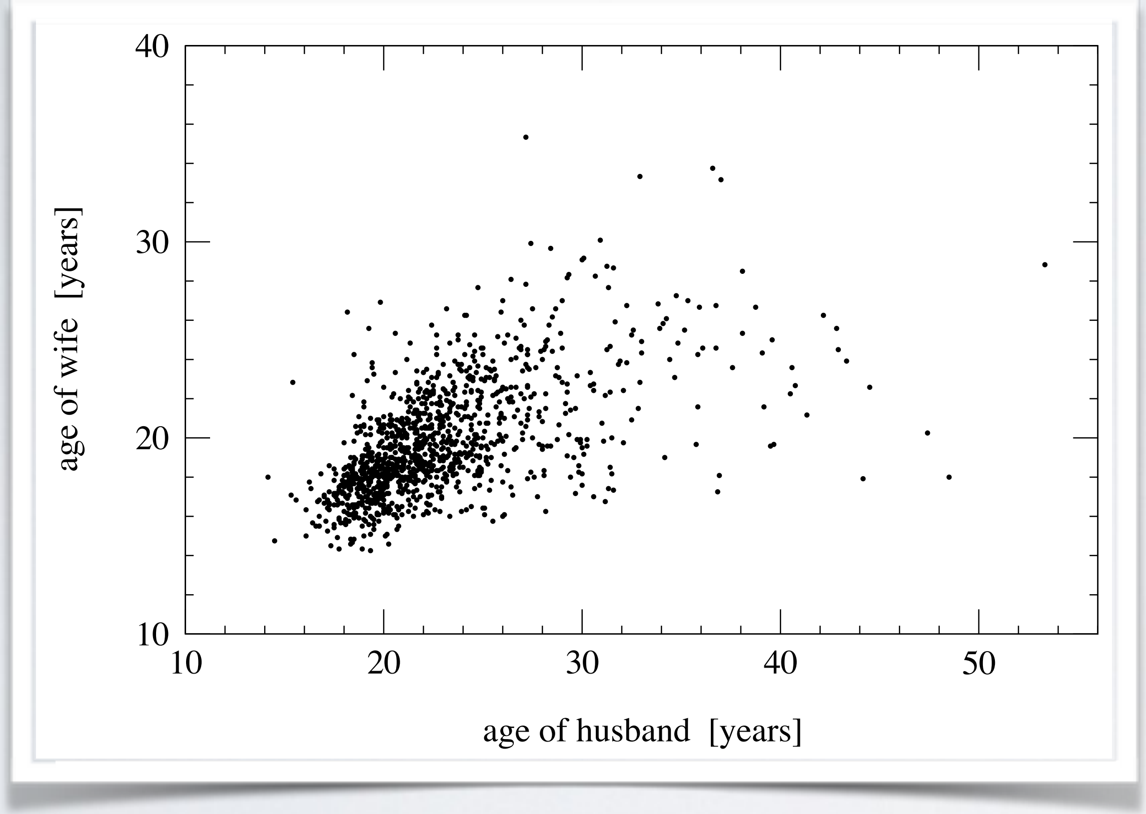
No assortative mixing : r=0 ($e_{ij} = a_i b_j$)

Perfectly assortative: r=1

Perfectly disassortative: -1<r<0

Homophily - Assortative mixing

To quantify homophily



Scalar properties

Pearson correlation coefficient of properties at both extremities of edges

e_{xy}: fraction of edges joining nodes with values x and y

$$\sum_{xy} e_{xy} = 1, \qquad \sum_{y} e_{xy} = a_x, \qquad \sum_{x} e_{xy} = b_y$$

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b},$$

with σ_a standard deviation of a_x

r=0, no assortative mixing,
r>0 assortative mixing,
r<0 disassortative mixing

Degree-degree correlation

- A particular type of application is the degree correlation:
  - Are *important nodes* connected to other important nodes with a higher probability than expected?
  - The degree can be used as any other scalar property

	network	type	size n	assortativity r	error σ_r
social	physics coauthorship	undirected	52 909	0.363	0.002
	biology coauthorship	undirected	1 520 251	0.127	0.0004
	mathematics coauthorship	undirected	253 339	0.120	0.002
	film actor collaborations	undirected	449 913	0.208	0.0002
	company directors	undirected	7 673	0.276	0.004
	student relationships	undirected	573	−0.029	0.037
	email address books	directed	16 881	0.092	0.004
technological	power grid	undirected	4 941	−0.003	0.013
	Internet	undirected	10 697	−0.189	0.002
	World-Wide Web	directed	269 504	−0.067	0.0002
	software dependencies	directed	3 162	−0.016	0.020
biological	protein interactions	undirected	2 115	−0.156	0.010
	metabolic network	undirected	765	−0.240	0.007
	neural network	directed	307	−0.226	0.016
	marine food web	directed	134	−0.263	0.037
	freshwater food web	directed	92	−0.326	0.031

M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

Rich-club coefficient

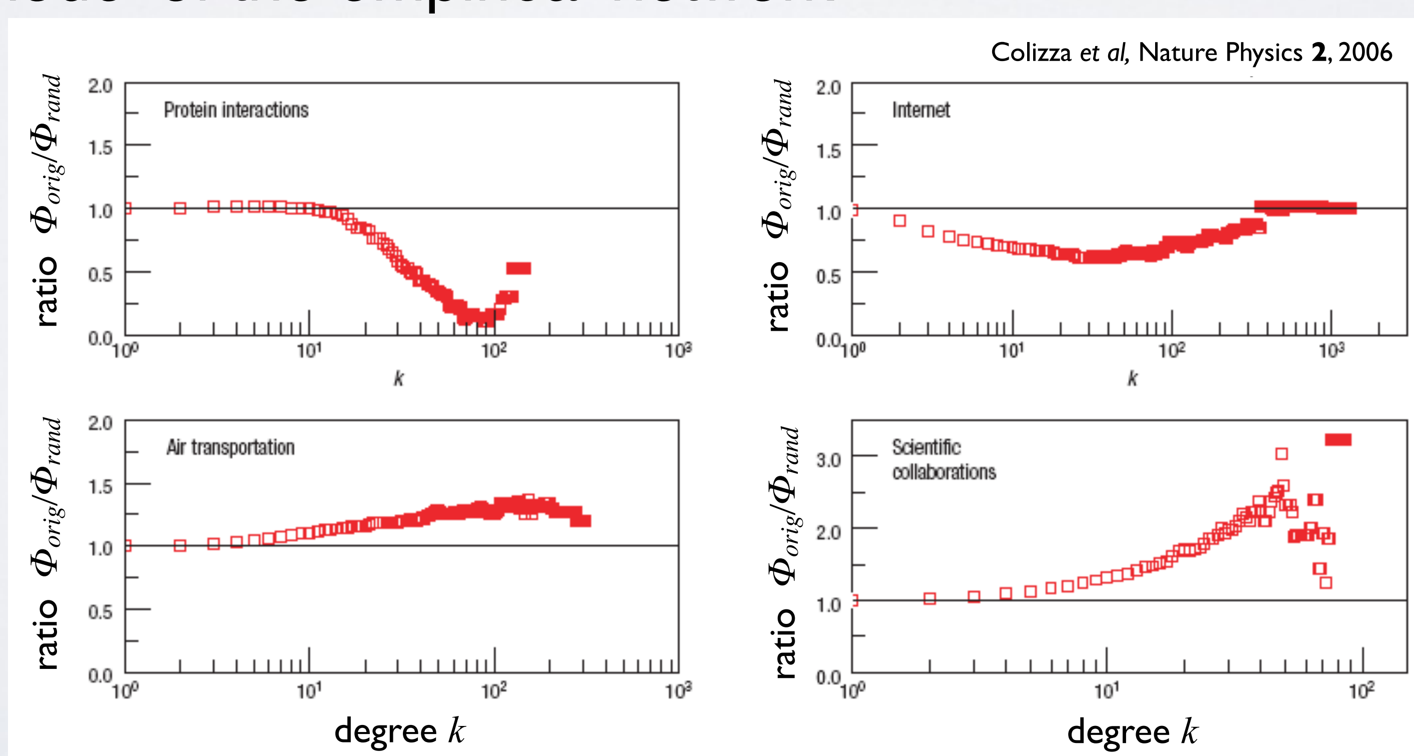
- How well connected are the well connected among themselves
- It is calculated on a list of node degree sorted in ascendant order as

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}$$

- $N_{>k}$ denotes the number of nodes with degree k or larger than k
- $E_{>k}$ measures the number of links between them
- Results are usually compared to random references
  - configuration model of equivalent synthetic network
  - configuration model of the empirical network

Algorithm

- rank nodes by degree
- remove nodes in an ascendant degree order
- measure the density of the remaining network

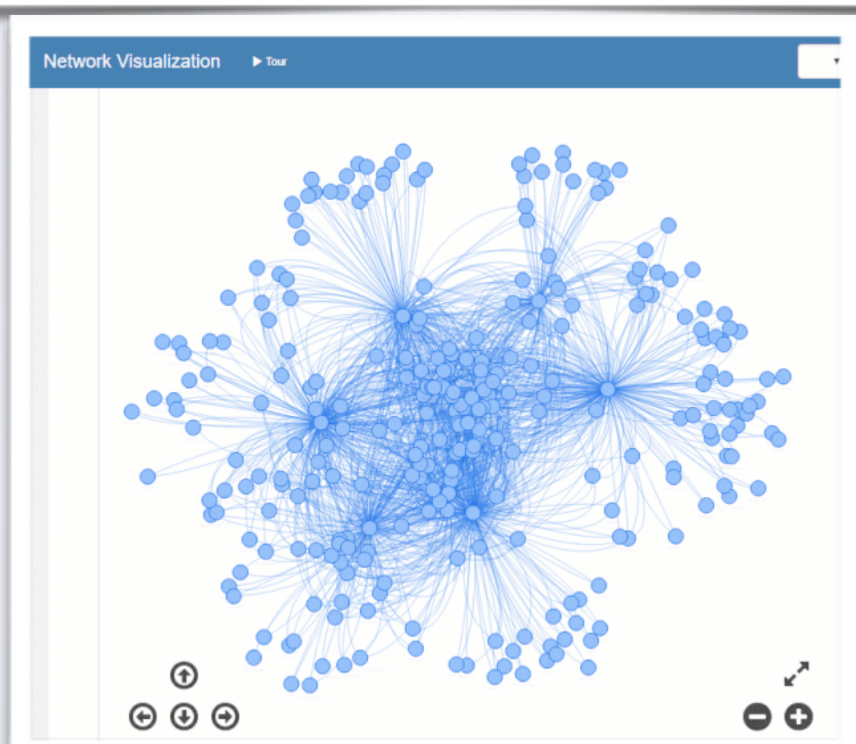


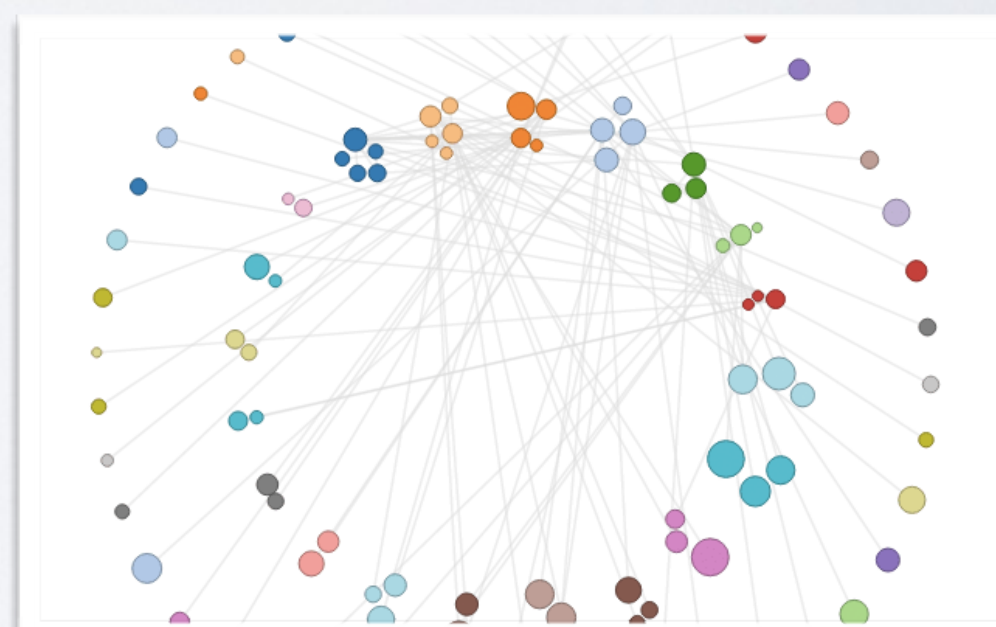
NETWORK VISUALISATION

NETWORK VISUALIZATION

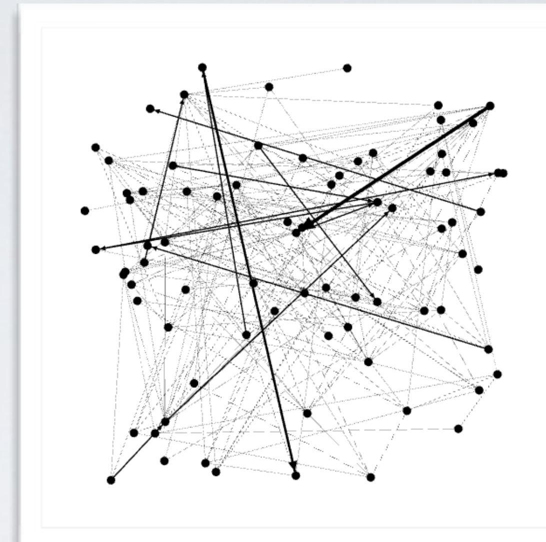
- How to interpret a network drawing?
- What does the position of nodes means?
- Can we draw conclusion from the drawing alone?

NETWORK VISUALIZATION

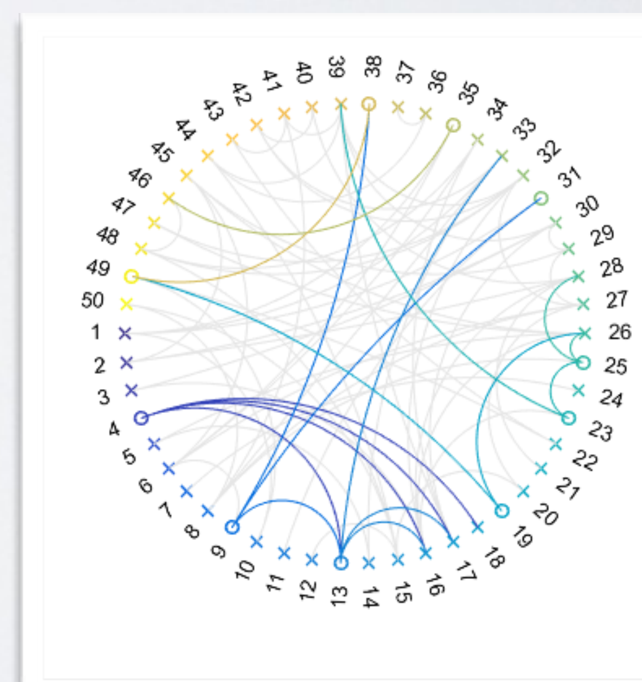




NETWORK VISUALIZATION

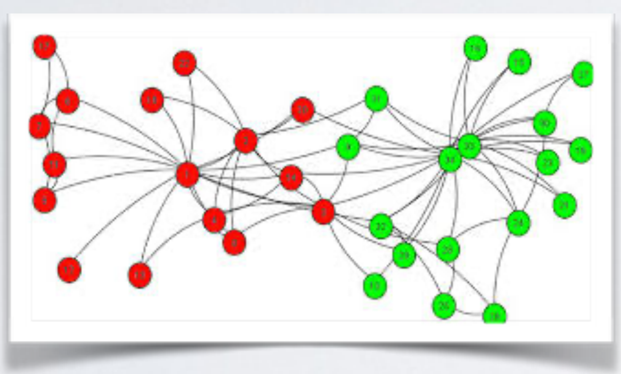


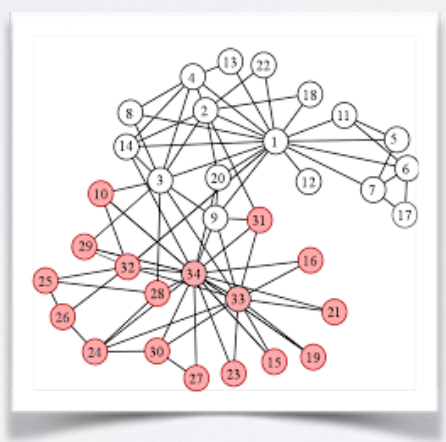
- Random layout
  - Assign random positions to nodes, draw edges
    - Useless for more than 5-6 nodes
- Geographical layout
  - The position of nodes is fixed apriori, often based on geographical location
  - Variant: position nodes on a circle based on a single, 1D property (age…)

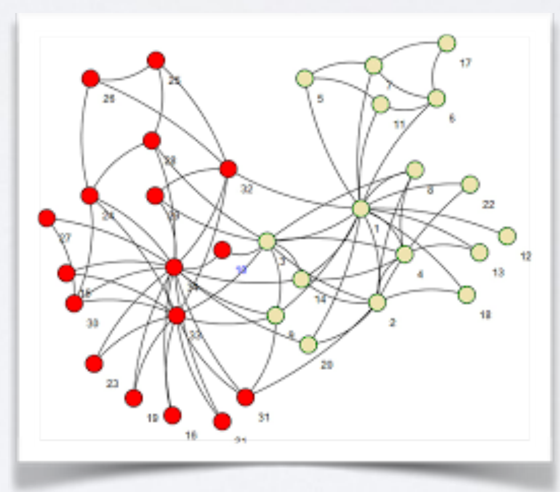


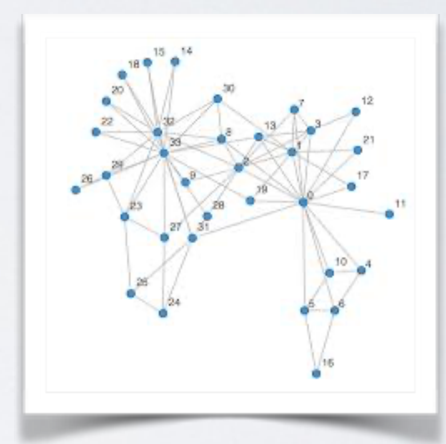
NETWORK VISUALIZATION

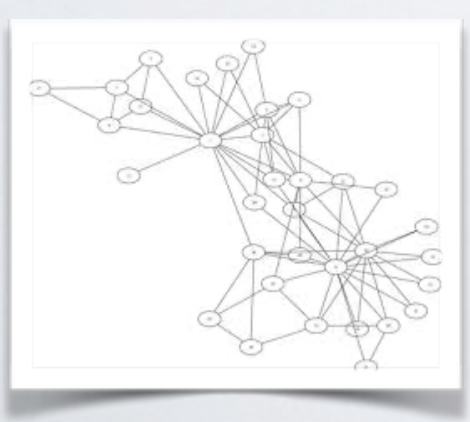
- Most commonly used: Automatic layout
  - Non deterministic
  - Tries to arrange nodes so that the network is easy to read and understand
    - Minimize edge crossings?
    - Most commonly, tries to put connected nodes close and unconnected nodes far

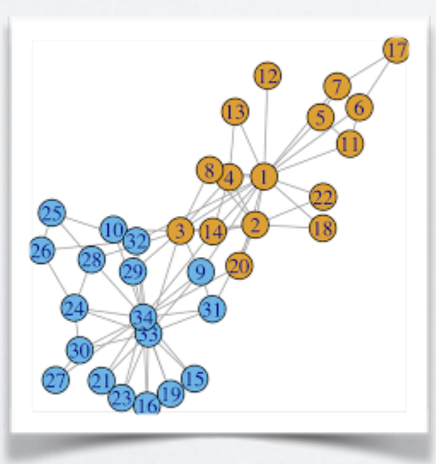






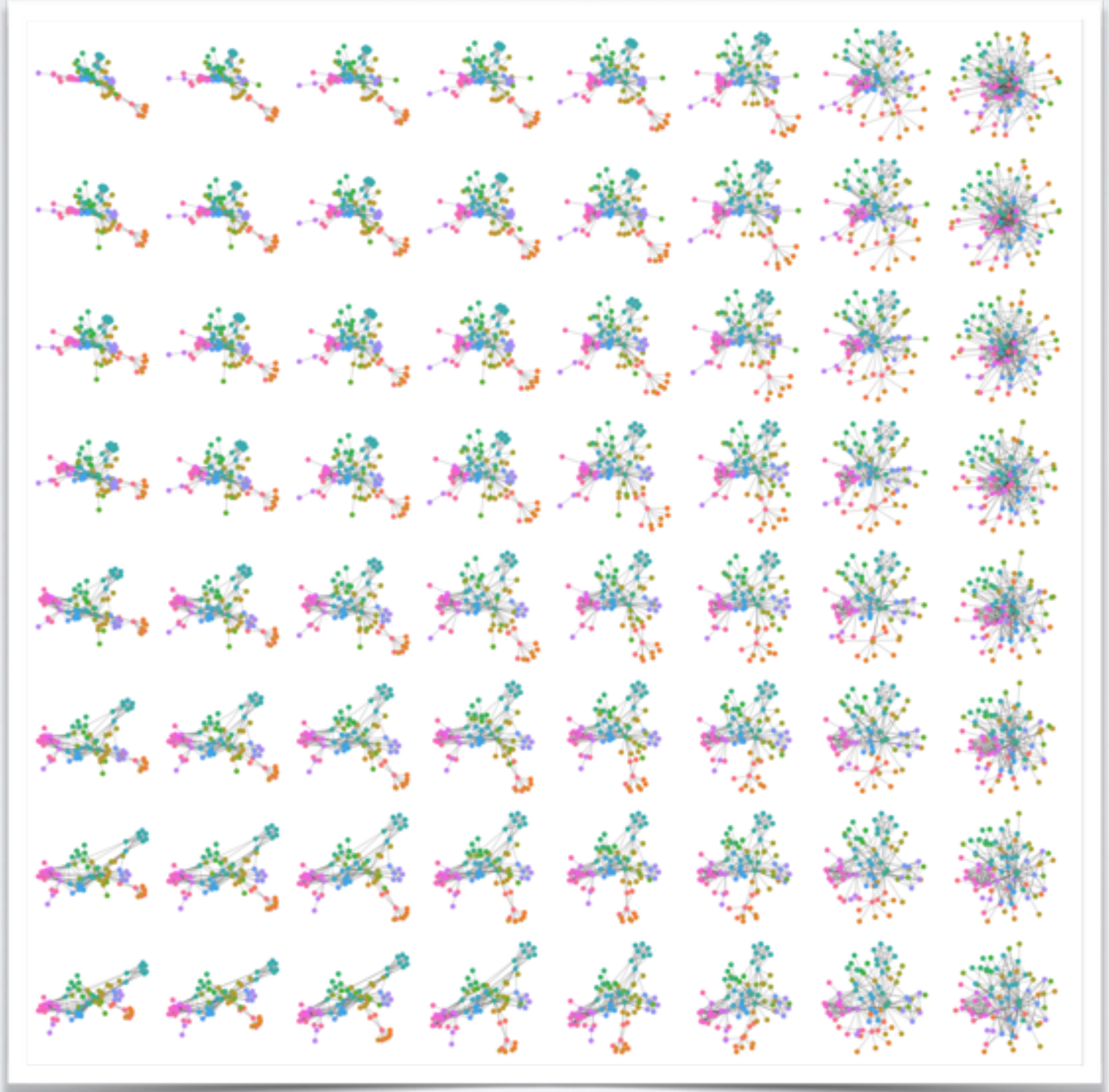






NETWORK VISUALIZATION

http://kwonoh.net/dgl/

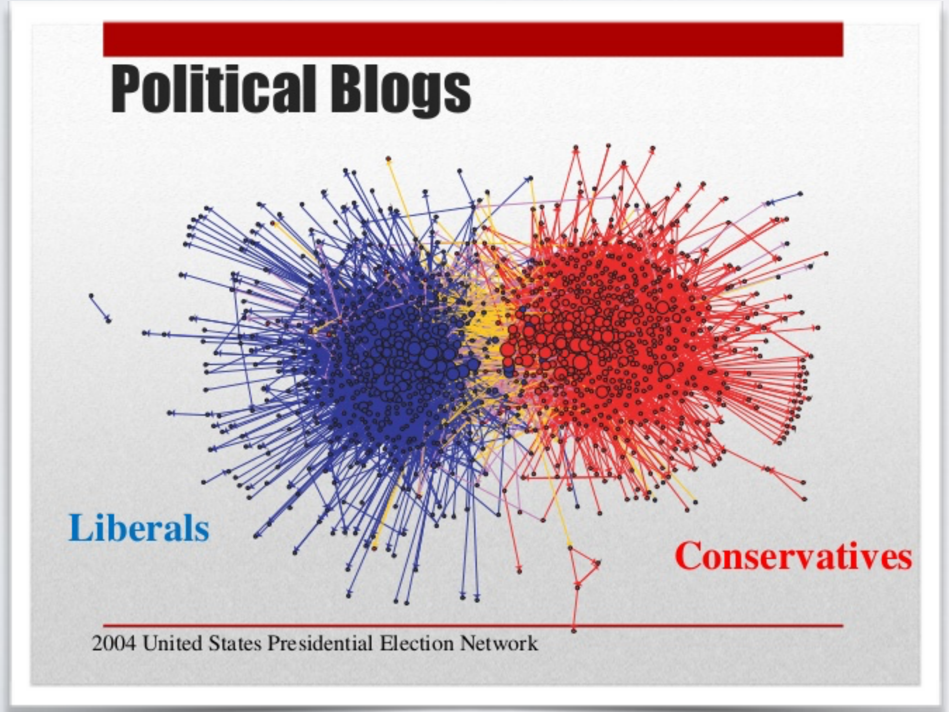


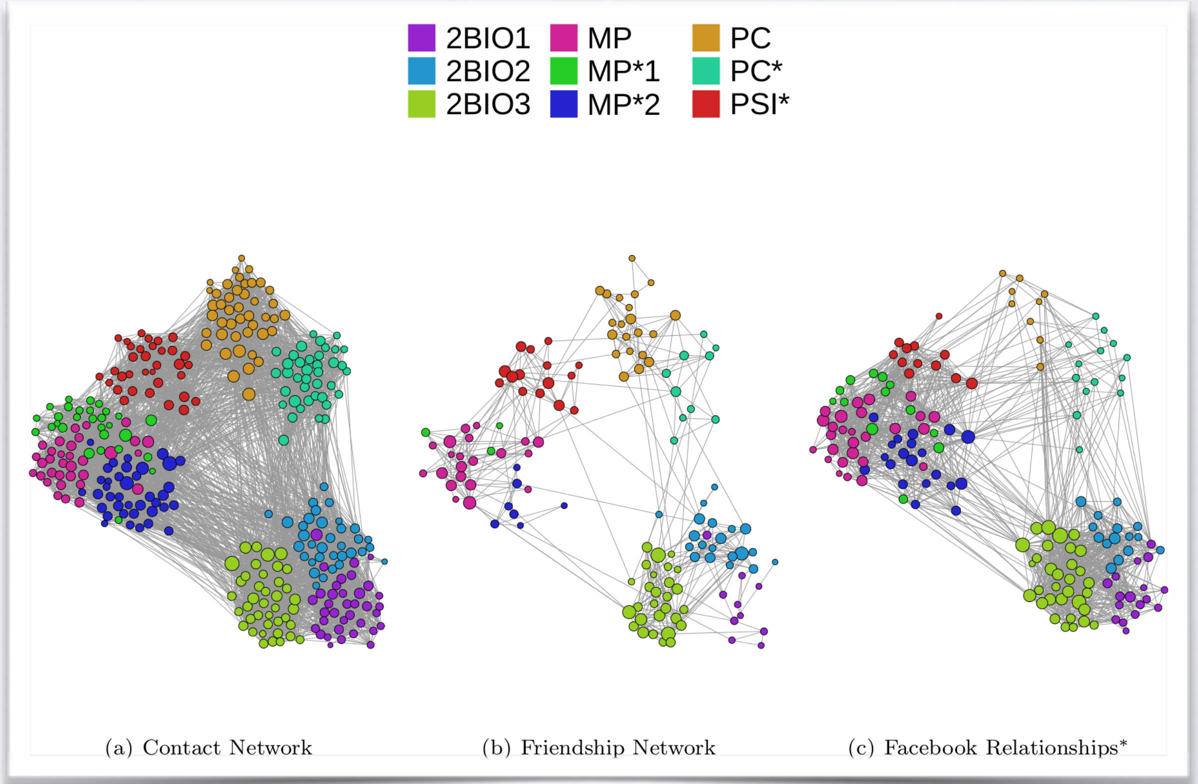
NETWORK VISUALIZATION

- Most common algorithms are variant of the **force directed** layout
  - Kamada-Kawai
  - Fruchterman-Reingold
  - …
- Force directed layout: a simple physical model
  - Repulsive forces between nodes
  - Edges are attracting forces
  - There are minimal (to avoid node overlap) and maximal (to avoid connected component drifting out of the figure) distances

NETWORK VISUALIZATION

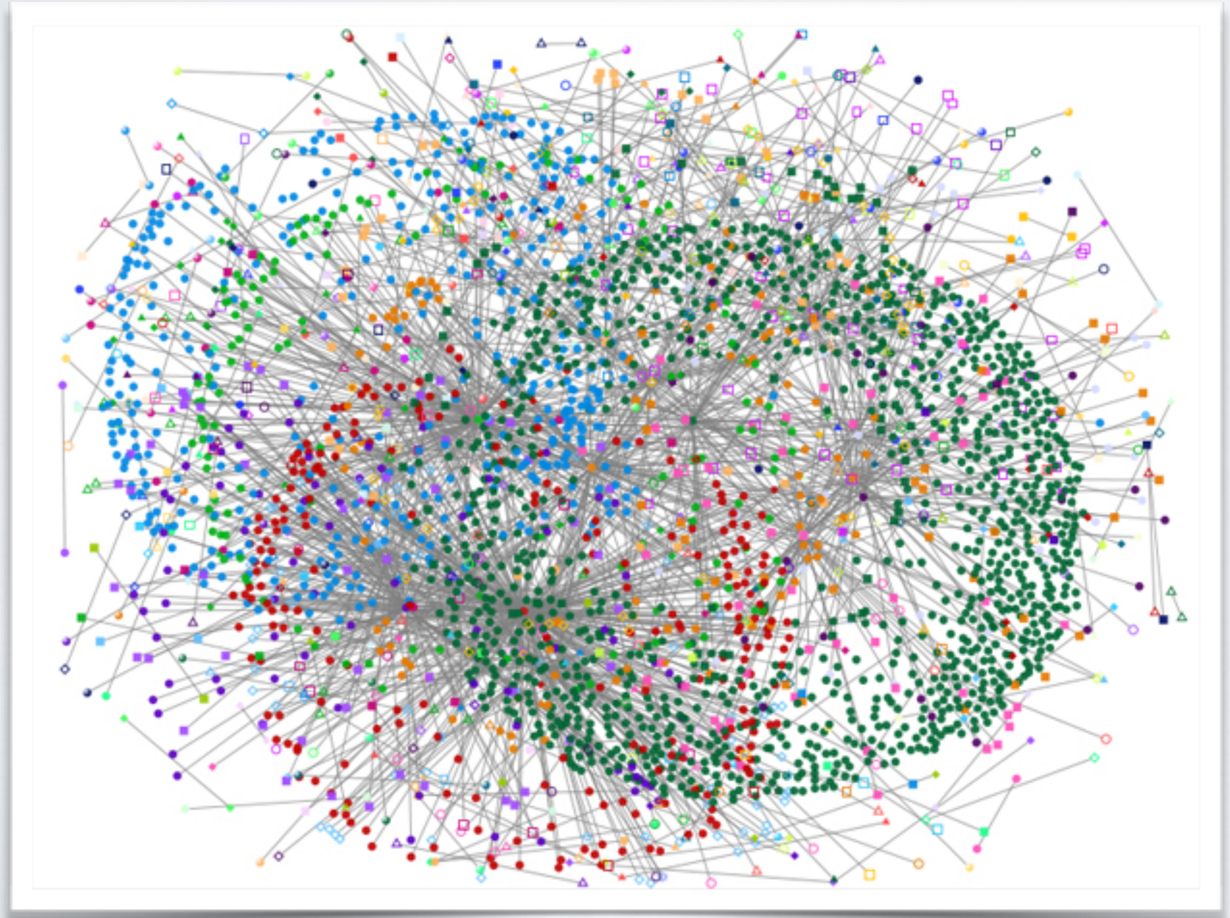
- Can we interpret a force layout?
  - Yes…

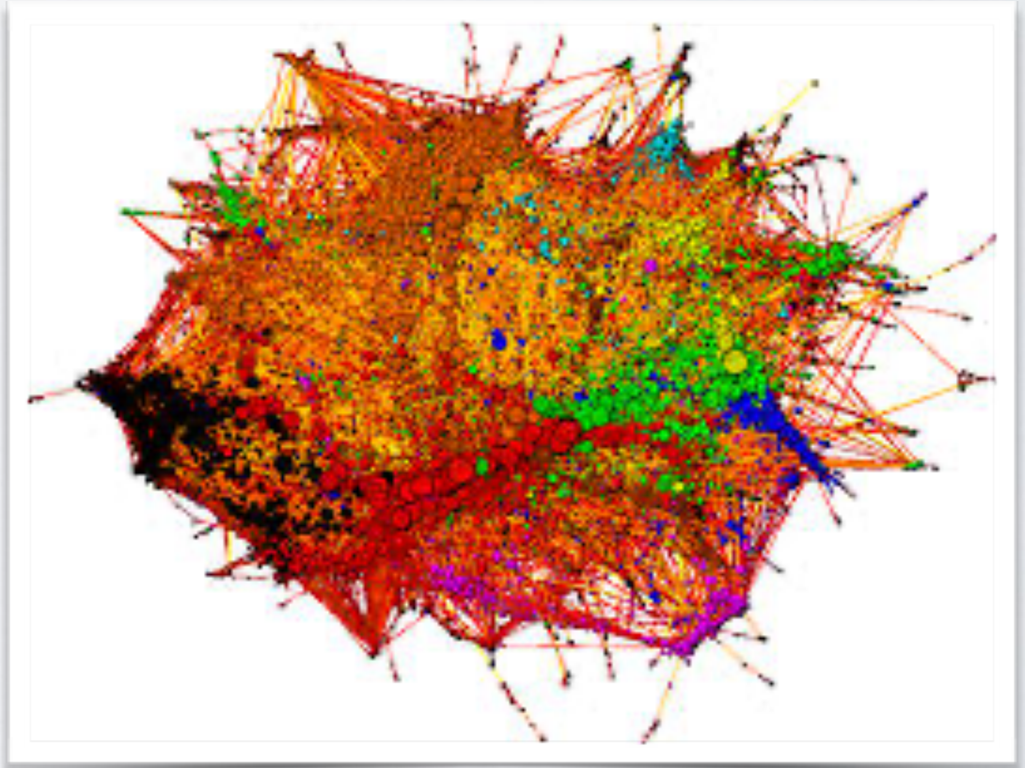




NETWORK VISUALIZATION

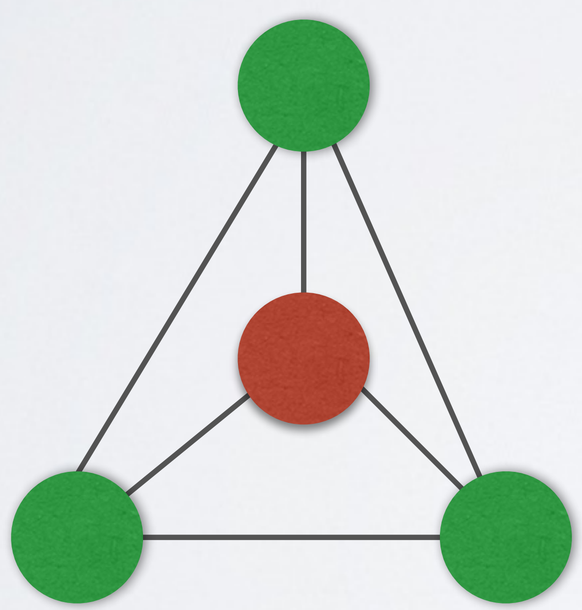
- Can we interpret a force layout?
  - Yes…
  - And no.

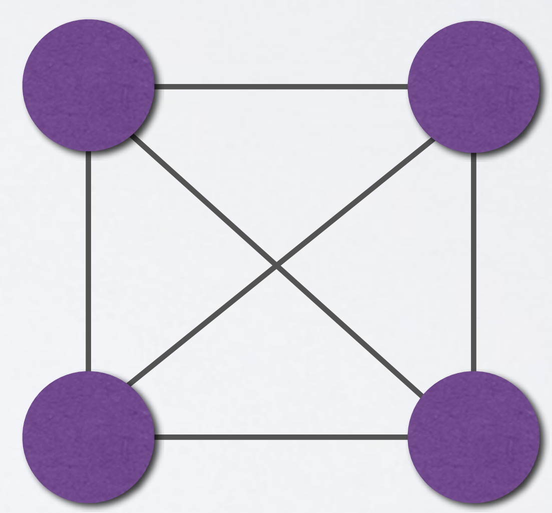




NETWORK VISUALIZATION

- Can we interpret a force layout?
  - Yes…
  - And no.





WHAT TO DO NOW

- http://cazabetremy.fr/Teaching/BitcoinNetwork.html
- Download the two provided networks. Choose one and load it with Gephi